5E-02

量子コンピューティングによる記事の組み合わせ最適化

木内美波 † 工藤和恵 † 中島寛人 ‡ 澤紀彦 ‡
お茶の水女子大学 † 日本経済新聞社 ‡

1 はじめに

現在、パソコンやスマホなどの電子機器でニュースを見る機会が増えた。ネットニュースは日々発行されており、日本経済新聞社が一日に発行してる記事は約900本である。大量の記事の中から、様々なジャンルのおすすめの記事のいくつかをピックアップしてユーザーに提供できれば、より効率的な情報の提供が可能になる。組み合わせ最適化を用いたプログラムでおすすめの記事の組み合わせを求めることをを目標とする。使用したソルバーは量子アニーリングマシンの D-Wave である [1]。プログラムの作成には組み合わせ最適化に関する先行研究を参考にした [2]。問題は QUBO (Quadratic Unconstrained Binary Optimization) 形式のハミルトニアン (目的関数) で表す。その最小値を求めることで最適な組み合わせが得られる。

2 設定

使用する記事のデータには発行された日時、タイトル、文字数、推薦度スコア (以下スコア)、ジャンルの情報が含まれている。この研究で求めたい最適な組み合わせとは、ジャンルの重複度と合計文字数の制約を満たし、できるだけ合計スコアが高い組み合わせのことを指す。また、記事の発行された時間が前日の18時から6時までを朝の記事、6時から12時までを昼の記事、12時から18時までを夕方の記事とし、それぞれの時間帯で最適な組み合わせを求める。

3 モデル

偏ったジャンルの記事ばかりが選ばれるのを避けるため、ジャンルの重複度の制約を設定した。合計文字数については、理想の文字数を朝は6000字、昼と夕を3000字とし、この理想の文字数との誤差を制約とした。

記事の組み合わせ最適化のモデルとして、次式のようなハミルトニアンを使用した。

$$H = w_1 H_1 + w_2 H_2 - H_3 \tag{1}$$

ここで、w_1, w_2 は正のパラメタである。三つの項からなるハミルトニアンで、設定した制約も項の中身に組み込んでいる。

一つ目の項は文字数に関する項で、次式で与える。

$$H_1 = \Big(\frac{1}{\alpha}\frac{1}{L^*}\big(\sum_i l_i x_i - L^*\big)\Big)^2 \tag{2}$$

ここで、x_i は二値変数であり、記事 i が選択された時に $x_i = 1$ とし、選ばれなかった場合 $x_i = 0$ とする。l_i は記事 i の文字数である。L^* は設定した理想の合計文字数である。合計文字数と理想の合計文字数との差を理想の合計文字数で割り、文字数の理想との誤差を表す。さらにこの値を α で割る。本研究では誤差10%を制約の範囲内とする。すなわち $\alpha = 0.1$ とする。よって実際の文字数の誤差がちょうど10%の時、H_1 の値が1になる。

二つ目の項は重複度に関係する項で、次式で与える。

$$H_2 = \frac{1}{\beta}\sum_{i<j} f_{i,j} x_i x_j \tag{3}$$

記事 i と記事 j が同じジャンルだった場合 $f_{i,j} = 1$、そうでない場合は $f_{i,j} = 0$ とする。β は重複度の制約の強さを表す。本研究では重複度3以下を制約の範囲内とする。よって重複度がちょうど3のとき $H_2 = 1$ となるように $\beta = 3$ とする。

三つ目の項はスコアに関する項で、次式で与える。

$$H_3 = \sum_i \frac{p_i - p_{\min}}{p_{\max} - p_{\min}} x_i \tag{4}$$

p_i は記事 i のスコア、$p_{\min}$ はスコアの最小値、$p_{\max}$ はスコアの最大値である。

式 (1) で表したハミルトニアンが最小の値を取るとき、最適な組み合わせを得られる可能性が高い。

4 方法

量子アニーリングで計算を行うために、元のデータの値をいくつか調整した。ジャンルなしの記事は予め

Combitional optimization of articles using quantum computing
†Minami Kiuchi, Kazue Kudo, Ochanomizu University
‡Hiroto Nakajima, Norihiko Sawa, Nikkei Inc.

除外した。複数のジャンルに分類されている記事は、記事の特性を最もよく捉えたジャンルを用いた。つまり全ての記事に必ず1つジャンルが与えられている。次に、6000字以上の記事についてはハミルトニアンに組み込む際に全て6000字として扱うことにした。文字数が6000字よりも多い記事は数は少なく、例外的な存在ながらも、結果に影響を与えるものが多いためである。

D-Waveは全結合問題を解く場合に扱える変数の数に限りがある。また、試行の結果、100以上の記事を一度に計算するのは難しいことが分かった。そのためプログラムに使用する記事をスコアが上位の記事30, 40, 50の3パターンに絞り込んだ。

D-Waveでは結果にムラがあったり、一回の実行でうまく答えが出ない場合もあるため、ハミルトニアンのパラメタを与えて、その最小化を1000回繰り返す。その解の中で制約を満たし、かつ最も高いスコアが得られる結果を出力する。

パラメタ w_1, w_2 は $0.1 \leq w_1 \leq 0.5$, $0.1 \leq w_2 \leq 0.5$ の範囲で設定した。

5 結果

古典的に計算した場合とD-Waveで計算した場合のスコアの違いを観察するために、シミュレーテッドアニーリング (以下SA) でも同様の計算を行った。図1は、SAで計算した場合のスコアのヒートマップである。横軸を w_1、縦軸を w_2 とし、各点の色がスコアを表している。データは3月17日の月曜の朝の記事で、計算にはそのうちのスコア上位の30個の記事を使用した。

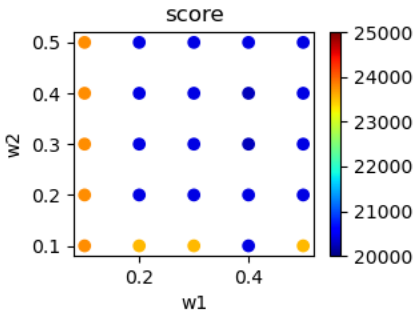

図1: SAで計算したスコアのヒートマップ

図2は、D-Waveの結果で作成したヒートマップである。使用した記事や条件は図1で用いたものと同様である。

図1を見ると、同じ色が近くに並んでいることが分かる。一方で図2では、色がバラバラに並んでいるように見える。また、SAより高いスコアの点もあること

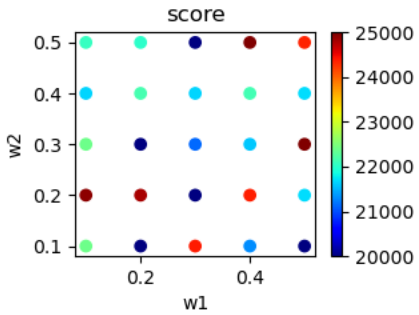

図2: D-Waveで計算したスコアのヒートマップ

が分かる。よってD-WaveはSAと比べて、同じ解に固定されず計算の度に多様な解を求められる特徴があることが分かった。よって、D-Waveではスコアが高い特定のパラメタの組み合わせを求めるよりも、パラメタの範囲を絞り、その中で高いスコアが出る組み合わせを都度選ぶやり方が望ましい。

6 まとめ

本研究では、量子アニーリングを利用して最適な記事の組み合わせを求めた。D-Waveでは一回のアニーリングにかかる時間は数十〜数百 μs と短い。短時間で多様な解を求めることが可能であり、その中にはSAでは手間と時間をかけないと得られないような高いスコアの解も含まれていた。よって高いスコアの組み合わせを効率よく求めたいという本研究の目的には、D-Waveを使用することに優位性があると言える。

今回は特定の日付のみで調整を行った。より一般的に使用できるようにするために違う週の同じ曜日のデータで実行する際の結果の違いを観察する必要がある。今後は、計算に使用する記事数が変わると結果の精度に影響するのかどうかを調べ、使用する上位記事数をさらに絞りこむ。また、量子アニーリングに関して、現在とは違う設定でも実行し、スコアが変化するかを試す。

参考文献

[1] D-Wave, 量子コンピューティング, https://dwavejapan.com/system/ (2021年12月17日アクセス).

[2] N. Nishimura, K. Tanahashi, K. Suganuma, M. J. Miyama, and M. Ohzeki, Item Listing Optimization for E-Commerce Websites Based on Diversity, Front. Comput. Sci. **1**, 2 (2019)